Using Data Grid Technology to Manage Distributed Data

Reagan W. Moore
University of California, San Diego
San Diego Supercomputer Center
moore@sdsc.edu
http://www.npaci.edu/DICE/

Generic Technology Evolution

- Community specification of intent
- Identification of common infrastructure across
  - Digital libraries
  - Data grids
  - Persistent archives
- Focus on scalability and robustness of unifying infrastructure

Storage Resource Broker (SRB)

*User communities brokered by SDSC instances of SRB***

As of 5/17/2002

Project Instance	*Data_size (in GB)*	*Count (files)*	*Comments*	*Funding Agency*
NPACI	1,972.00	1,083,230	NPACI Users	NSF/PACI
Digsky	17,800.00	5,139,249	2MASS,DPOSS,NVO	NSF/ITR
DigEmbryo	433.00	31,629	Visible Embryo	NLM
HyperLter	158.00	3,596	HyperSpectral Images	NSF/NPACI (ESS)
Hayden	6,800.00	41,391	FlyThrough for Planetarium	AMNH/Hayden
Portal	33.00	5,485	Grid Portal	NSF/NPACI
SLAC	514.00	77,168	Protein Crystallography	NSF/NPACI (Alpha)
NARA	7.00	2,455	Archival Documents	NARA
SIO Exp	19.20	383	SIO Explorer Documents	NSF/NSDL
ADL	0.00	6	ADEPT Digital Library	NSF/DLI2
TRA	5.80	92	Classroom Videos	NSF/NPACI (EOT)
DTF	239.00	1,766	DTF users	NSF/TCS
AfCS	27.00	4,007	Cell Signalling Images/Docs	NIH
TOTAL	28,008.00	6,390,457		
	28 TB	6.4 million		

** Does not cover data brokered by SRB spaces administered outside SDSC.
Does not cover databases; covers only files stored in file systems and archival storage systems

Topics

- Data management systems
  - Data Grids, Digital Libraries, Persistent Archives
- Common data management technology
  - Logical name space, storage abstraction
- Collection federation
  - Knowledge management systems

Digital Libraries

- Provide services on the data collection
  - Ingestion, loading of attribute values
  - Extensibility, definition of new attributes
  - Discovery, queries on attributes
  - Browsing, hierarchical listing
  - Presentation, formatting specified data models
- Communities
  - Digital library
  - Global Grid Forum, Databases and the Grid working group
  - OMG, Common Warehouse Metamodel

Data Grids

- Manage data in a distributed environment
  - Logical name space, provide global identifier
  - Data access, storage system abstraction
  - Replication, disaster back up
  - Uniform access, common API across file systems, archives, and databases
  - Single sign-on, authenticate across administration domains
- Communities
  - Global Grid Forum, data grids
  - Discipline specific data management systems

Persistent Archives

- Manage technology evolution
  - Storage system abstraction, support data migration across storage systems
  - Information repository abstraction, support catalog migration to new databases
  - Logical name space, support global persistent identifier
- Communities
  - Persistent archive community
  - Global Grid Forum, Persistent archive working group

Common Capabilities

- Logical name space
  - Registration of digital entities
- Storage repository abstraction
  - Operations used to manipulate data in a storage system
- Information repository abstraction
  - Operations used to manipulate a catalog in a database

Data Grid
(Storage Resource Broker)

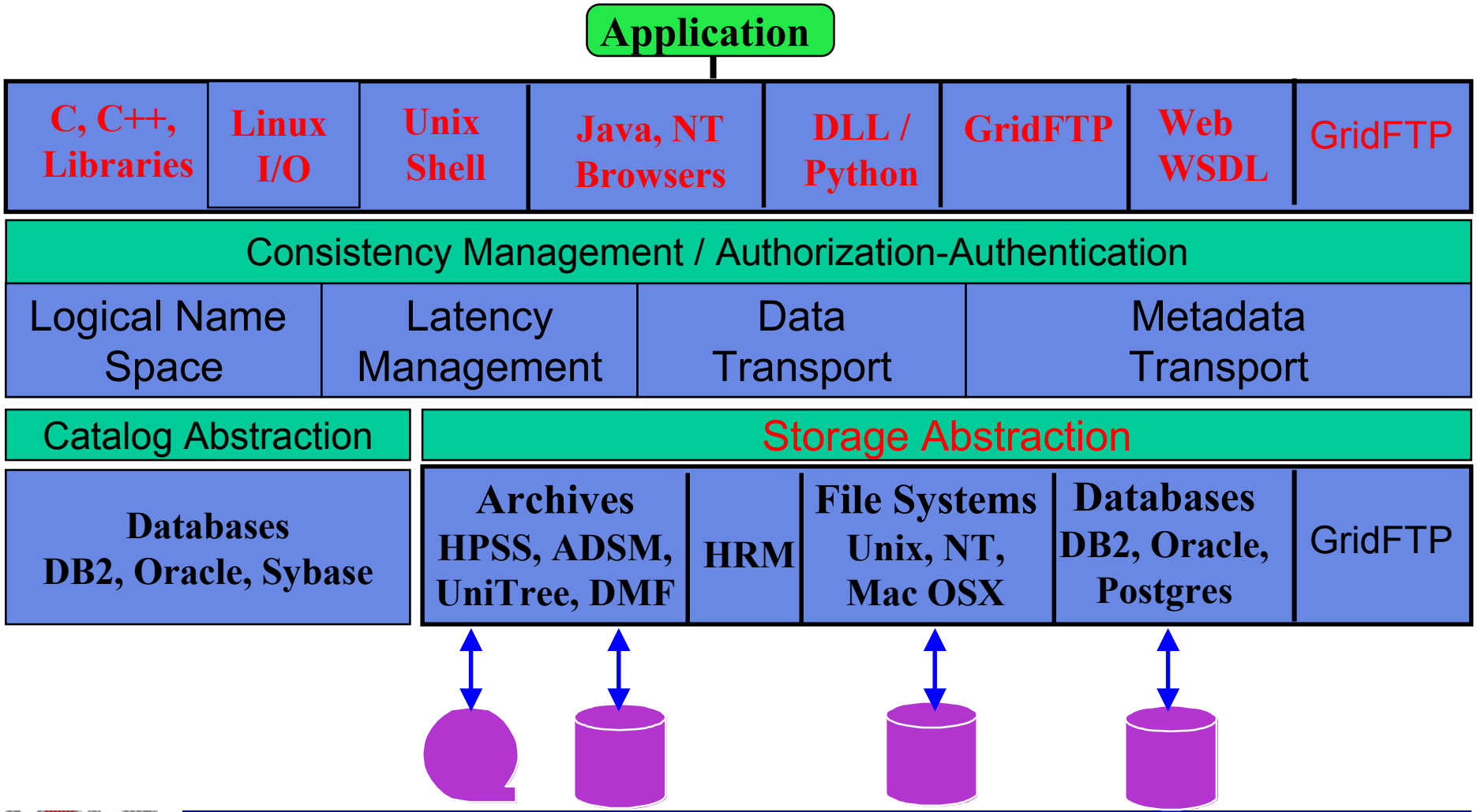
- Integration of collection-based management of digital entities, with
  - Remote data access through storage system abstraction
  - Catalog access through information repository abstraction
  - Automation through collection-owned data

Storage Abstraction

- Provide common access semantics
  - Archival storage systems
  - File systems
  - Databases
- Support Unix file system operations
  - Map from the interface preferred by your application to the interfaces required by legacy storage systems
- Support database interactions
  - Map from information repository abstraction to database commands

SDSC Storage Resource Broker & Meta-data Catalog

Storage Abstraction



Logical Name Space (Data Grid Transparencies)

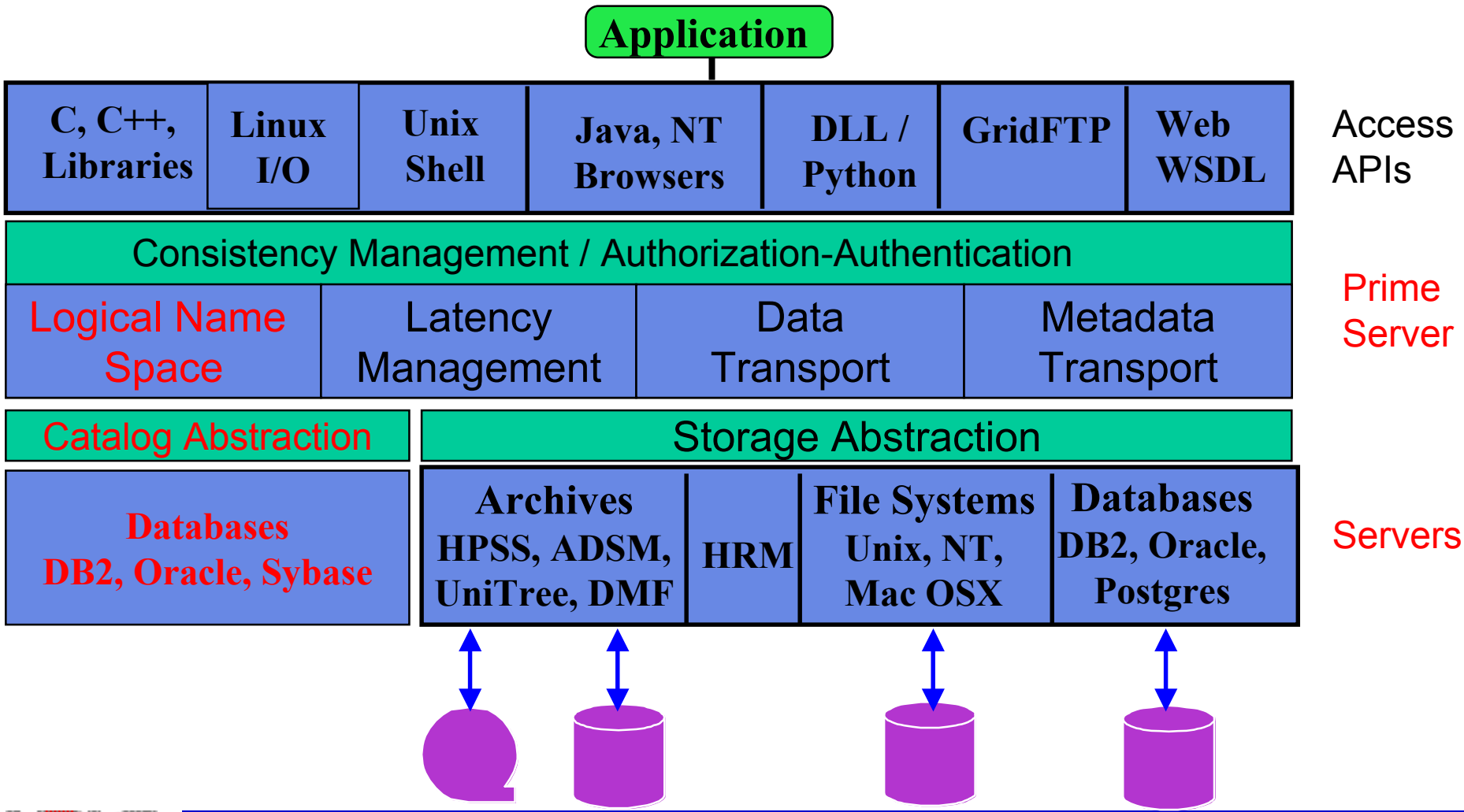
- Naming transparency - find a data set without knowing its name
  - Map from attributes to a global file name
- Location transparency - access a data set without knowing where it is
  - Map from global file name to local file name
- Access transparency - access a data set without knowing the type of storage system
  - Federated client-server architecture

Logical Name Space Operations

- Replication
  - One to many mapping from logical name to physical name
- Containers
  - Mapping from logical name to location in a physical container
- Shadow links
  - Registration of user owned data into the collection

SDSC Storage Resource Broker & Meta-data Catalog

Logical Name Space



Digital Entities

- Digital entities are “images of reality”, made of
  - Data, the bits (zeros and ones) put on a storage system
  - Information, the attributes used to assign semantic meaning to the data
  - Knowledge, the semantic and structural relationships described by a data model
- Every digital entity requires information and knowledge to correctly interpret and display

Types of Digital Entities

- Files
  - Physical files in the collection ID space
  - Shadow links to files in your user ID space
- Directories
  - Shadow links to directories in your user ID space
- Databases
  - Shadow links to tables
  - SQL command strings
- URLs

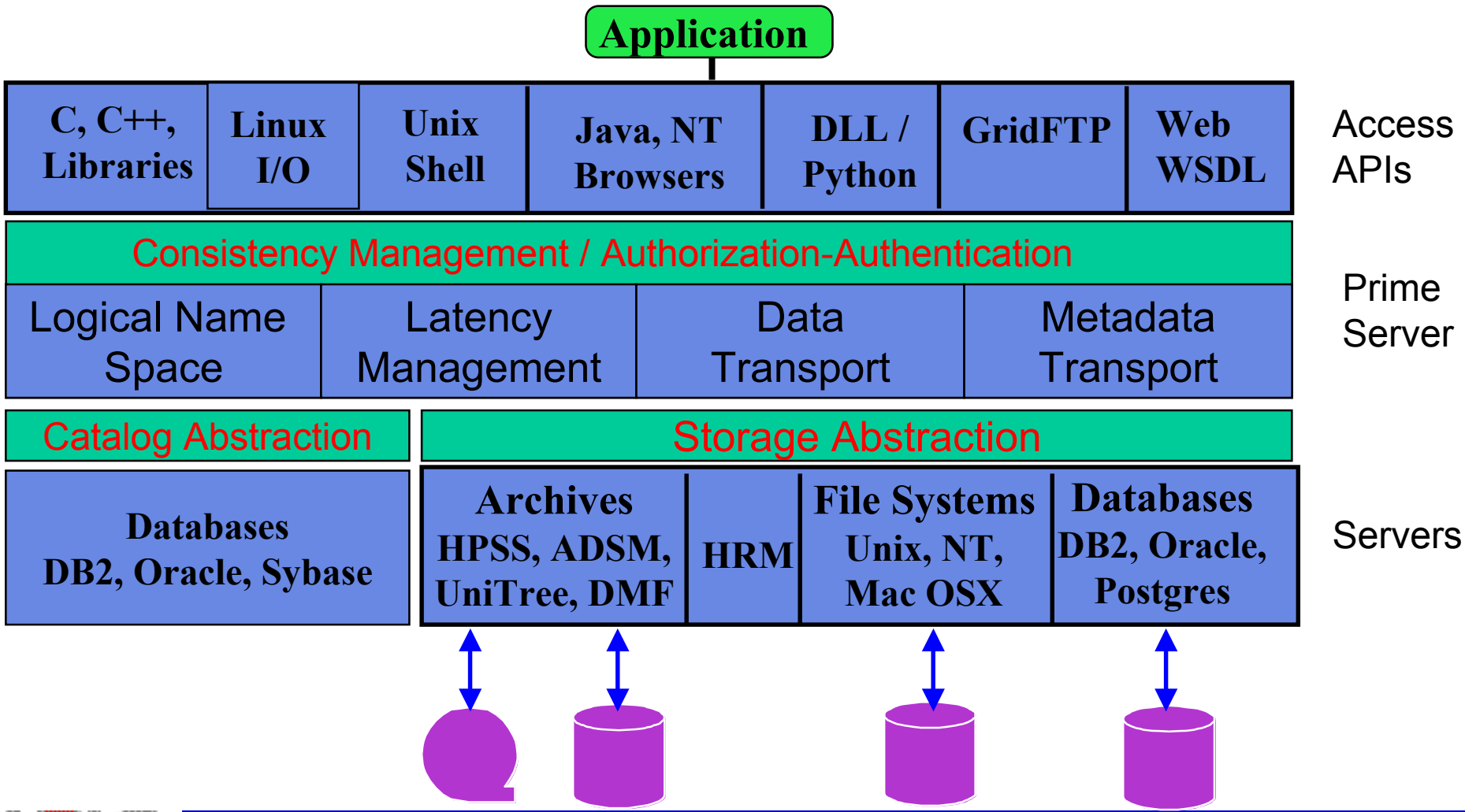
Preservation

(Similar requirements to a data grid)

- Name transparency
  - Find a file by attributes (map from attributes to global name)
- Location transparency
  - Access a file by a global identifier (map from global to local file name)
- Access transparency
  - Use same API to access data in archive or file cache
- Authenticity
  - Disaster recovery, replicate data across storage systems
  - Audit and process management

SDSC Storage Resource Broker & Meta-data Catalog

Preservation



Convergence of Technologies

- Data grids as basis for distributed data management
  - Federation of distributed resources
  - Creation of logical name space to automate discovery
- Digital libraries
  - Discovery based on attributes
  - Hierarchical collection management
  - Extensible schema through information repository abstraction
- Persistent archives
  - Data replication
  - Persistence management

Data Naming Ontologies

Concept space	Discipline concepts
Collection	Discipline attributes
Data grid	Global Identifier
Archive / file systems	Local file name
Data model	Attributes that describe data structure

Differentiating between Data, Information, and Knowledge

- **Data**
  - Digital object
  - Objects are streams of bits
- **Information**
  - Any tagged data, which is treated as an attribute.
  - Attributes may be tagged data within the digital object, or tagged data that is associated with the digital object
- **Knowledge**
  - Relationships between attributes
  - Relationships can be procedural/temporal, structural/spatial, logical/semantic, functional

Knowledge Creation Roadmap

- Knowledge syntax (consensus)
  - RDF, XMI, Topic Map
- Knowledge management (recursive operations)
  - Oracle parallel database
- Knowledge manipulation (spatial/procedural rules)
  - Generation of inference rules and mapping to data models
- Knowledge generation (scalable inference engine)
  - Application of inference rules in inference engine

Knowledge Based Data Grid Roadmap